NLP

Week2

Word representation

V = [a, aaron, …, zulu, <UNK>]

1-hot representation

Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
$\begin{bmatrix}0\\0\\0\\0\\\vdots\\1\\\vdots\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\\vdots\\1\\\vdots\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix}$	$\begin{bmatrix}0\\\vdots\\1\\\vdots\\0\\0\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\0\\\vdots\\1\\\vdots\\0\end{bmatrix}$

I want a glass of orange _____.

I want a glass of apple_____.

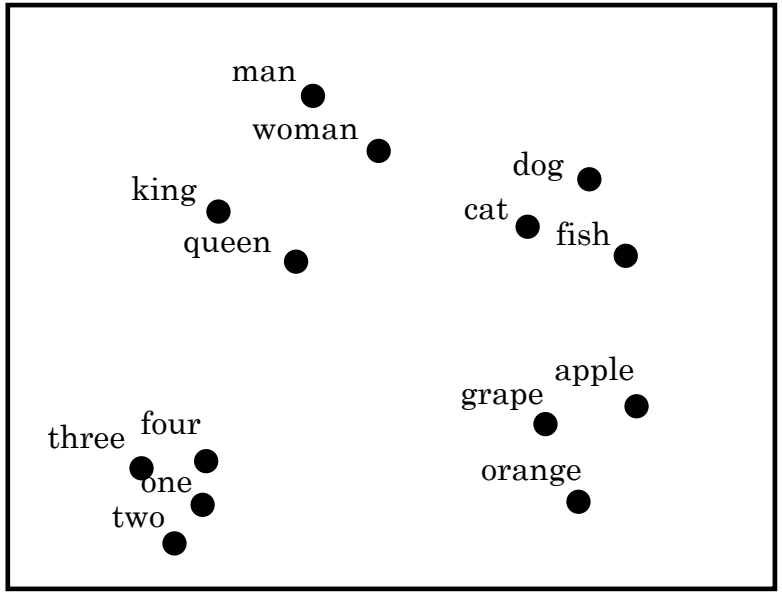
Featurized representation: word embedding

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97

I want a glass of orange _____.

I want a glass of apple_____.

Visualizing word embeddings



Transfer learning and word embeddings

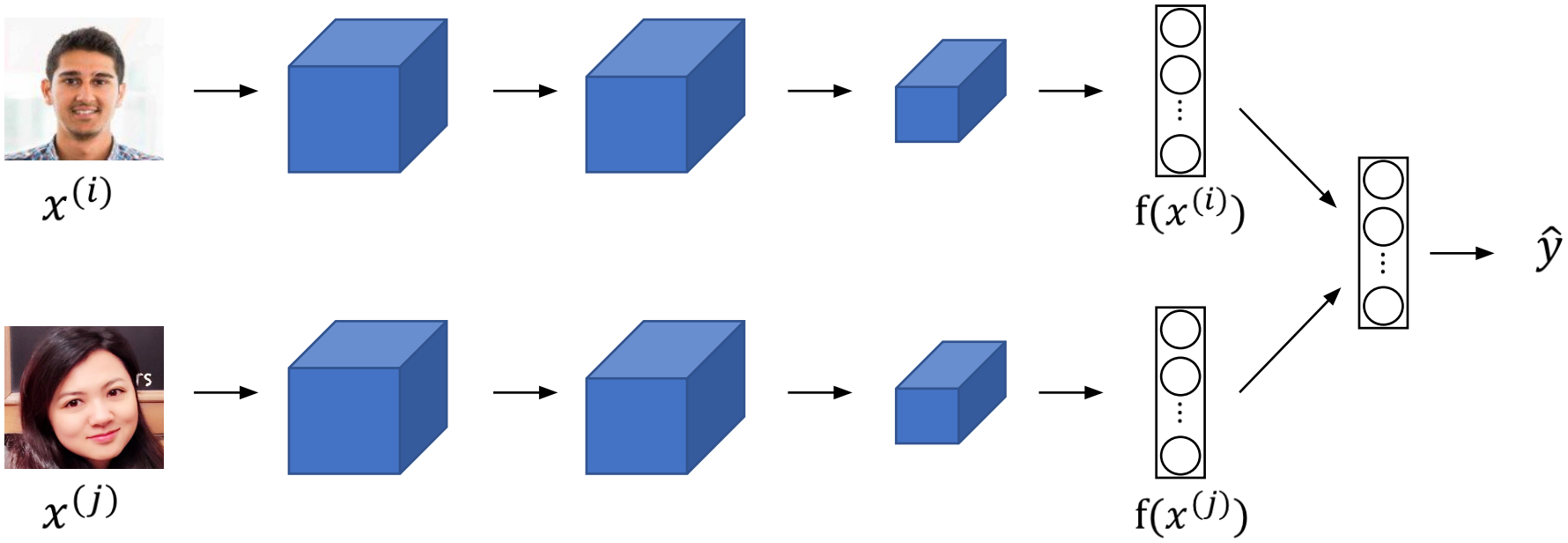
1. Learn word embeddings from large text corpus. (1-100B words)

   (Or download pre-trained embedding online.)

2. Transfer embedding to new task with smaller training set. (say, 100k words)

3. Optional: Continue to finetune the word embeddings with new data.

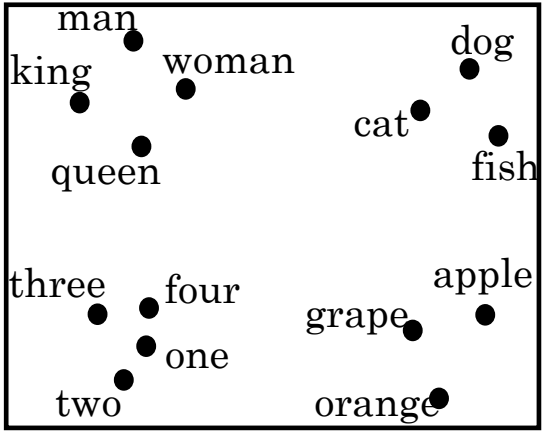
Relation to face encoding



Analogies

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

Analogies using word vectors



$$e_{man} - e_{woman} \approx e_{king} - e_?$$

Cosine similarity

$$sim(e_w, e_{king} - e_{man} + e_{woman})$$

Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia

Embedding Matrix

	a (0)	Apple (456)	...	...	....	Orange (6257)	....	Queen (7157)	<UNK> (9999)
Gender	....	0.00	....	....	....	0.01	....	0.97	....
Royal	....	-0.01	....	....	....	0.00	....	0.95	....
Age	....	0.03	....	....	....	-0.02	....	0.69	....
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....
Food	....	0.95	....	....	....	0.97	....	0.01	....

Orange (6257)

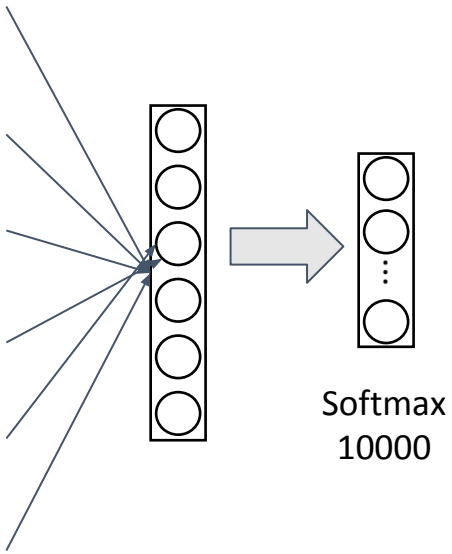
$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$E \,.\, O_{6257} = e_{6257} = \text{embedding for word j}$$

Neural language model

I	want	a	glass	of	orange	_____.
4343	9665	1	3852	6163	6257	

I	o_{4343}	$\longrightarrow$	E	$\longrightarrow$	e_{4343}
want	o_{9665}	$\longrightarrow$	E	$\longrightarrow$	e_{9665}
a	o_1	$\longrightarrow$	E	$\longrightarrow$	e_1
glass	o_{3852}	$\longrightarrow$	E	$\longrightarrow$	e_{3852}
of	o_{6163}	$\longrightarrow$	E	$\longrightarrow$	e_{6163}
orange	o_{6257}	$\longrightarrow$	E	$\longrightarrow$	e_{6257}



Other context/target pairs

I want a glass of orange juice to go along with my cereal.

Target

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Context: Last 4 words.

4 words on left & right

Last 1 word

Nearby 1 word

Skip-grams

I want a glass of orange juice to go along with my cereal.

Context	Target
orange	juice
orange	glass
orange	my

Model

Vocab size = 10,000k

Context (c) Orange$_{[6257]}$ $\longrightarrow$ Target (t) Juice$_{[4834]}$

$$O_c \longrightarrow E \longrightarrow e_c \longrightarrow \text{softmax} \longrightarrow \hat{Y}$$

Softmax $P(t|c) = \dfrac{e^{\theta_t e_c}}{\sum_{j=1}^{10000} e^{\theta_j e_c}}$

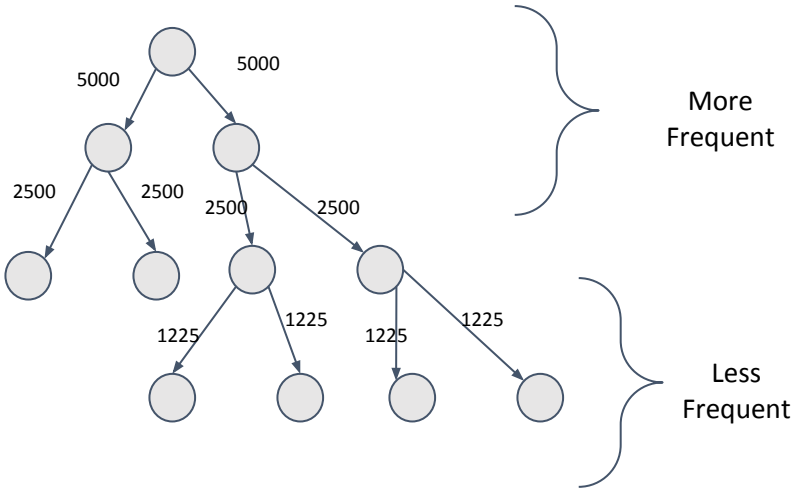
$$Y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

θ_t = Parameters associated with the target t

Problems with softmax classification

$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

Hierarchical Softmax



How to sample the context c?

Don't Select Stop words for context selection

Defining a new learning problem - Negative Sampling

I want a glass of orange juice to go along with my cereal.

context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

For 1 +ve sample select K negative sample

K can be 5 - 20 for smaller data sets
K can be 2 - 5 for larger data sets

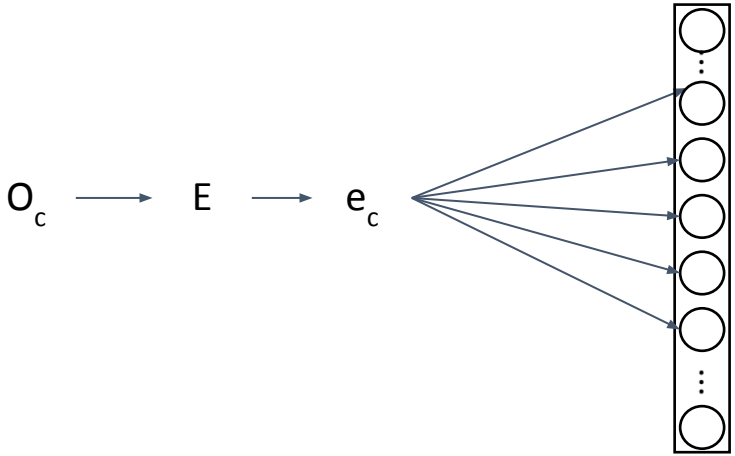
Model - Negative Sampling

Softmax: $$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

Binary Classification problem so using sigmoid

$$P(1/0 \mid c,t) = \sigma(\theta_t e_c)$$

context	word	target?
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0



For a given training set, instead of training all the 10000 binary logistics units, we will be training on (K+1) units, thus speeding up the classification

GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal.

Context

Target

X_{ij} = Number of time i_t appears in the context of j_c

X_{ji} = Number of time j_c appears in the context of i_t

$$X_{ij} = X_{ji}$$

Model

$$\text{minimize } \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij})\left(\theta_i^T e_j + b_i + b_j' - \log X_{ij}\right)^2$$

Weighting Term

$(\theta_t e_c)$

$f(X_{ij}) = 0$ if $X_{ij} = 0$

$X_{ij} = X_{ji}$

So, θ_i and e_j are symmetric

$\theta_w^{final} = (e_w + \theta_w) / 2$